Automated Bundle Pagination Using Machine Learning

Alessandro Torrisi, Robert Bevan, Katie Atkinson, Danushka Bollegala, and Frans Coenen
Department of Computer Science, University of Liverpool, Liverpool, L69 3BX, UK
{torrisi,robert.bevan,k.m.atkinson,danushka.bollegala,coenen}@liverpool.ac.uk

ABSTRACT

Coherent division of legal document bundles, whether this is done in the context of court bundles, briefs or some other application, is a time consuming and challenging task. We propose an approach whereby this process can be automated. Two variations are considered. The first addresses the scenario where the topic labelling is pre-defined and adopts a supervised learning approach. The second addresses the scenario where the topic labelling, for whatever reason, is not specified in advance and adopts an unsupervised learning approach. This paper reports on an investigation of both mechanisms using accident claims bundles. The evaluation results indicate that the proposed approaches can be successfully applied to divide legal document bundles.

CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**; **Information extraction**.

ACM Reference Format:
Alessandro Torrisi, Robert Bevan, Katie Atkinson, Danushka Bollegala, and Frans Coenen. 2019. Automated Bundle Pagination Using Machine Learning. In *Seventeenth International Conference on Artificial Intelligence and Law (ICAIL '19), June 17–21, 2019, Montreal, QC, Canada.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3322640.3326726

1 INTRODUCTION

We tackle the task of dividing a document into discrete pages. In information processing terms, this process involves segmenting an electronic document, or set of electronic documents, into a set of individual pages ready for printing, a process embedded into many software systems such as word processors and text editors. In the legal domain, we often need to structure and order legal document bundles ("collections") according to subject/topic. Essentially, we derive a sorted list of information and evidence relevant to a legal case, whilst also including an additional index, summary and chronology.

Pagination is a challenging problem to apply in practice because of the varied content and size of the document bundles to be paginated. Regardless of whether court bundles, briefs or any other form of legal document bundle is under consideration, the document collections are typically unstructured, are comprised mostly of scanned documents and can run to several thousand pages. Pagination is therefore a time consuming task for any legal concern.

This paper presents a machine learning approach to automated processing of document bundles. The idea is that given a document bundle, the pages are paginated according to k subjects/topics. The subjects can be predefined in terms of a set of k labels, $L = \{c_1, c_2, \ldots, c_k\}$, or can be automatically identified from within the content of the document. Pagination using predefined labels is essentially a **supervised learning** classification problem, whereas pagination without predefined labels is essentially an **unsupervised clustering** problem. Both approaches are considered in this paper. Whatever the case, the idea is that each page in a document bundle is assigned a class label so that the bundle is divided according to topic. Such a document representation enables users to easily detect the boundaries between different kinds of information in the bundle. It also allows users to arrange the bundle according to the topics (classes).

On completion of the process, each page is tagged with a reference date and a short text summary. A reference date is typically related either to: (i) when the report was typed or (ii) when the event described occurred. Reference dates were extracted from text by exploiting the presence of specific keywords appearing in the proximity of parsed dates (such as *referral*, *date reported*, *ref*, *date/inserted* and *attendance*). However, there will be pages that do not contain any reference date. This is likely to happen when a block of related pages include a reference date only in the first page. In this case, all the pages not containing any dates were assumed to be the continuation of the last reference date found in the text. In some cases it may be the case that indexing information is included in the page header or footer, in which case this can be utilised for grouping pages. Finally, the text summary, for each page, was automatically extracted using the TextRank summarisation algorithm [15, 16].

The remainder of this paper is organised as follows. A review of related work is presented in Section 2. Section 3 then presents the supervised pagination mechanism where the labels are prespecified, whereas Section 4 presents the unsupervised mechanism where the labels are derived by the system. The evaluation of the proposed pagination approaches is presented in Section 5 together with some discussion. Finally, some conclusions are presented in Section 6.

2 PREVIOUS WORK

Text classification is the process of automatically labelling textual documents with a predetermined set of categories (classes). It is a popular task in Machine Learning [11, 12]. In the legal domain, text classification can be used to organise large data collections in a structured manner. A good overview of standard natural language processing tasks and techniques, including text preprocessing, classification and clustering, is given in [1]. Examples of where the

ICAIL '19, June 17–21, 2019, Montreal, QC, Canada

ACM ISBN 978-1-4503-6754-7/19/06…$15.00
https://doi.org/10.1145/3322640.3326726

utility of text mining has been explored in the context of the legal domain can be found in [19] and [4].

The work reported in [19] was directed at an investigation of the application of text classification with respect to a corpus of 131,830 court rulings, from the French Supreme Court, up until 2016; although not in the context of pagination. The Bag of Words (BoW) feature vector representation, based on unigrams and bigrams, was used. Three scenarios were considered: the categorisation of cases according to legal area, the prediction of the outcome of a case in the data set and the prediction of the period when the case was adjudicated on. Different classification models were used for each, although in each case the model was built using an ensemble of multiple Support Vector Machines [9]. Eight classes defining the legal areas were used, six and eight classes defining outcomes (two sets of experiments) and seven time interval classes. The reported $F1$ measures of 98%, 96% and 87% for the three classifiers respectively indicate a successful utilisation of text classification techniques in the legal domain.

In [4] the use of text classification was considered in the context of court dockets as used in the United States' legal system. Three scenarios were experimented with: the detection of docket errors, the matching of orders with motions and, as in the case of [19], outcome prediction. For the docket error detection, context information was added. Four categories of docket error were considered; it was argued that automatically detecting these types of errors enhanced the quality control procedures required prior to the submission of a legal bundle to court. The matching of orders with motions was treated as an information retrieval task given that it was likely that there would be a high correlation between the two. Motions and orders were represented using Term Frequency - Inverse Document Frequency (TF-IDF) and compared using the cosine similarity. Outcome prediction was conducted in a binary manner using a binary classification model trained using n-gram frequency vectors where n ranged from 1 to 4 (a minimum frequency of eight was used). For the evaluation, a 10-fold cross-validation was adopted together with a Support Vector Machine classifier [9]. The evaluation demonstrated that a good outcome prediction model ($F1 = 95\%$) could be learnt without involving any human intervention.

The work presented in [19] and [4] was not directed at pagination, although the work did indicate the potential for using text classification techniques for pagination purposes. To the best knowledge of the authors, there is no previous work directed at the pagination problem using text classification techniques, as proposed in this paper. However, there are examples where an unsupervised (clustering) approach has been applied to the pagination problem. Examples can be found in [8] and [20].

In [8] the authors reported on the outcomes from a series of clustering experiments considering large heterogeneous law firm collections of legal documents (such as law reports). They consider both hard clustering (a candidate document can only be assigned to a single cluster) and soft clustering (a candidate document can be assigned to multiple clusters) solutions. The hard clustering, that is of interest with respect to the work presented in this paper, was conducted using the well-known k-means clustering algorithm and the TF-IDF feature vector representation also used in [4] for classifying US legal system dockets. To evaluate the outcomes, paralegal researchers were asked to assess the quality of the resulting clustering. They rated the best clustering results achieved to be both topically coherent and useful to legal practitioners.

3 PAGINATION USING PREDEFINED (FIXED) LABELS

Given a legal document bundle to be paginated, in many cases the relevant legal teams are aware of the topics that they wish to be highlighted with respect to the pagination task. This knowledge comes from experience. This is particularly the case where legal firms are working in specific litigation domains, for example insurance litigation, where the legal team "know what they are looking for". The proposed mechanism is to build a classifier covering the identified set of topics, expressed in terms of a set of k classes $\{c_1, c_2, \ldots, c_k\}$. To build such a classifier we require training data. In the case of legal firms, where pagination is a regular activity, a rich repository of examples is typically available. For the evaluation presented in Section 5, medical claims data was used that had been previously manually paginated; a time consuming process.

Before any classification model can be built, the documents must be represented in some manner. For the proposed mechanism, each page of a selected bundle is represented by a feature vector representation, because this is the standard representation used with respect to most supervised and unsupervised learning models. Thus each page in each bundle is represented by a vector of features $V = \{v_1, \ldots, v_n\}$. Six feature vector representations were considered with respect to the work presented here, three founded on the Bag-of-Words (BoW) representation and three founded on the concept of Topic Modelling using LDA [3]: (i) Standard BoW, (ii) Normalised BoW, (iii) TF-IDF BoW, (iv) LDA generated using a standard BoW input, (v) LDA generated using a Normalised BoW input and (vi) LDA generated using a TF-IDF BoW input. A word in this context is a word n-gram (either a uni-gram, bi-gram and tri-gram).

Prior to generating the individual representations, the bundles are first pre-processed by applying stemming and stop word removal to the text. Punctuation symbols and other noisy characters, introduced as a result of the Optical Character Recognition (OCR) tool used when scanning bundles, are also removed. This is followed by word pruning where very frequently occurring words and rare words; which, by definition, will not be good discriminators of class, are removed from the collection. A frequently occurring word is considered to be one that appears in 80% or more of the pages in the bundle, whilst a rare word is considered to be one that appears in 20% or less of the pages in the bundle. The remaining words are then used to generate the required representations.

The BoW representation is the simplest. Using the BoW representation each page in the bundle is represented in terms of a feature vector of length n where n is the number of n-grams considered. For the Standard BoW the feature vector is defined in terms of a frequency histogram of the selected n-grams, one histogram (feature vector) per page. For the Normalised BoW representation the Euclidean norm $||\vec{V}||$ was used to give normalised frequency values, values between 0.0 and 1.0.

For the TF-IDF BoW representation, as the name suggests, TF-IDF values were used. TF-IDF is a common measure used in text mining applications [17] where, given a key word, its Term Frequency

(TF) in the current page is divided by the (Inverse) Document Frequency (IDF) across the bundle; the effect is to decrease the weight for commonly used terms and increase the weight for uncommon terms.

Topic modelling is a statistical technique for identifying "topics" in a collection of documents (pages within a bundle). LDA is a widely used topic modelling mechanism [3]. When using this mechanism, a document is represented as a mixture of topics that are present in the bundle. In the context of the proposed pagination system, topics are sets of n-grams (as defined above). Topics were extracted from text through a Gibbs sampling-based approach [10]. At first, each n-gram in the bundle is randomly associated to one of the k topics of interest; $k = 50$ was used with respect to the evaluation presented later in this paper. To improve the random assignment, an iterative approach was followed to optimise the product between the two probabilities $p_1(t|d)$ and $p_2(w|t)$ given as follows:

- $p_1(t|d)$: the proportion of words in page d that are assigned to topic t.
- $p_2(w|t)$: proportion of assignments to topic t, over all pages in the bundle, that come from n-gram w.

During each iteration, each word was reassigned, to the same or a new topic, with probability $p_1 \times p_2$. After a large number of iterations, the algorithm reaches a state where topic assignments are stable. Three different versions of LDA were considered, each founded on one of the BoW representations considered (see above).

Experiments were conducted using a number of classification techniques: Naïve Bayes [21], Logistic Regression [13] and Random Forest [5]. The results are presented in Section 5.

4 PAGINATION USING UNSPECIFIED LABELS

Where appropriate training data is not available, or the purpose of the pagination is exploratory in nature, in other words, the legal team does not know in advance what they are looking for, pagination using unspecified labels is required. In other words, we wish to cluster the bundle into k topics/classes. Again there are a range of clustering algorithms available but the simplest, and that adopted with respect to the evaluation presented later in Section 5, is k-means clustering [14]. The challenge here is on deciding the number of clusters (k) to be generated. Two potential mechanisms for determining the optimal clustering granularity are the Elbow method [2] and Average Silhouette analysis [18].

The idea of the Elbow method is to generate a set of cluster configurations using a range of values for k. For each cluster configuration, the Within Cluster Sum of Squares (WCSS) score is calculated. This provides a measure of how far the points inside a cluster are from their centroid. A line chart is then generated by plotting the calculated WCSS scores as a function of the number of clusters. The Elbow method selects a value for k in such a way that adding another cluster will not provide a better modelling of the data. In other words, the WCSS scores tend to decrease as k is increased. The location of an "elbow" in the line plot is generally considered to be the point where there is an abrupt decrease in the WCSS score. The corresponding k value is then selected as the appropriate number of clusters.

Silhouette analysis provides a measure of how similar each point in a cluster is compared to other clusters in a given cluster configuration. The Silhouette coefficient s for a point i is calculated by considering the mean distance between i and all other points in a cluster compared to mean distance between i and the points in the nearest neighbour cluster. Thus:

$$s_i = \frac{(b - a)}{max\,(a, b)} \tag{1}$$

where b is the mean distance between i and all the points in the nearest neighbour cluster, and a is the mean intra-cluster distance. Each silhouette coefficient ranges from -1 to $+1$, where a high value indicates that the point i is highly correlated with the current cluster. This is equivalent to saying that the resulting clusters are well defined with intra-cluster variances greater than inter-cluster similarities. The average silhouette score calculated across all the points quantifies the quality of a clustering configuration. Following this criterion, one should select the value of k which provides the maximum average silhouette score. With respect to the work presented in this paper, the Elbow method and Silhouette analysis were combined; the Elbow method to derive a value for k and the Silhouette analysis to confirm the selection.

Once the clusters had been generated, a second challenge was to assign labels to these clusters. Contrarily to the supervised case where labels are specified by the user, clustering provides only a way of grouping pages with similar contents. To provide clusters with explicative labels, the most frequent keywords appearing in each cluster were matched with a pre-built domain specific lexicon that included words labels. Experiments were conducted using the same six representations as those used for fixed label pagination. The results are presented in Section 5.

5 EVALUATION

To evaluate the fixed label and unspecified label automated bundle pagination process, medical record bundles were used of the form that might be used in accident claims litigation, as described in Sub-section 5.1 below. Experiments were conducted using the six different representations described in Section 3. The advantage of the fixed label, classification, approach is that there are well established methods for determining the effectiveness of such systems where the predicted class value can be compared to the known (ground truth) value [7]. For the fixed label pagination, three classification models were considered: Naïve Bayes, Logistic Regression and Random Forest. The accuracy results are given in Sub-section 5.2. For the unspecified label pagination approach, cluster separation and cohesion were used to measure the effectiveness of the approach; the accuracy results are given in Sub-section 5.3.

5.1 Data Set

An accident claims data set was used for the evaluation comprising 70 bundles that had been previously paginated, ranging in size from 17 to 6934 pages with an average size of 1858 pages[1]. For the fixed label pagination evaluation the label set $L = \{$*"blood test"*,

[1]Work is progressing on generating much larger training/test data sets, further categories will also be included. This dataset contains sensitive information which prohibits us from making it publicly available at this stage.

Average Cross Validation									
Representation	Naïve Bayes			Logistic Regression			Random Forest		
	Precis.	Recall	F1	Precis.	Recall	F1	Precis.	Recall	F1
Standard BoW	93.36%	92.99%	92.99%	**95.72%**	**95.66%**	**95.64%**	**97.55%**	**97.49%**	**97.5%**
Norm. BoW	**94.49%**	**94.33%**	**94.3%**	93.72%	93.49%	93.46%	**97.56%**	**97.49%**	**97.5%**
TF-IDF BoW	94.45%	**94.33%**	**94.3%**	95.09%	94.82%	94.79%	**97.55%**	**97.49%**	**97.49%**
LDA_{BoW}	92.1%	91.83%	91.72%	92.16%	91.83%	91.73%	90.58%	90.49%	90.28%
$LDA_{Norm\ BoW}$	81.22%	76.99%	74.05%	80.67%	79.16%	76.55%	77.1%	77.33%	76.82%
$LDA_{TF-IDF\ BoW}$	78.12%	76.33%	74.27%	79.22%	77.83%	75.56%	73.01%	72.33%	71.99%

Table (1) Average Classification performance (%).

"consent form", *"consultation note"*, *"gp record"*} was considered. For each category, 100 different sample pages were manually selected. For the fixed label pagination, a 5-fold Cross Validation (5CV) analysis was performed with each fold including 70% of the data as training set. For the unspecified label pagination, a bundle composed of 120 pages (30 pages in each category) was considered.

5.2 Fixed Label Pagination

For the evaluation of the fixed label pagination the metrics used were *Precision*, *Recall* and the *F1* measure. *Recall* gives us an indicator of a classifier's performance with respect to false negatives, while *Precision* gives an indicator of a classifier's performance with respect to false positives. The *F1* score is the harmonic average of the *Precision* and *Recall*. In the context of paginating document bundles, users are interested in overall performance, hence the *F1* measure is a good summarising measure.

Table 1 gives the average cross validation results obtained, in which it can be seen that the BoW representations produced the best *F*1 measures. LDA provides the best F1 measure only in the second validation set using Standard BoW and Naïve Bayes. The best results were produced by the Random Forest classifier coupled with a BoW representation (see line 1 in Table 1). A breakdown of the results obtained using the the Random Forest classifier is given in Table 2 with respect to the second validation set for which the best results were achieved. Table 2 indicates that there is a good quality of prediction over the four categories of pages. However, inspection of the test data indicated that there was a case where a *"consultation note"* page was associated with a *"gp record"* label. This is partly because of the nature of GP records which include information about a patient's health history, such as diagnoses, medicines, tests, allergies, immunisations and treatment plans. A consultation note represents correspondence between primary care physicians and specialists regarding a patient's condition. Terms like *"dear doctor"*, *"Yours sincerely"*,*"consultant"*,*"Re:"* are typically included in a consultation note. In this particular case, none of these were included in the consultation note. In addition, a list of medications was included in the note which led the classifier to associate it to *"gp record"*. A similar outcome was observed for the other validation sets where a consultation note was included in *"blood test"* and *"page record"* pages. The classifier associated the label *"consultation note"* to both pages.

5.3 Unspecified Label Pagination

For the evaluation of the unspecified label pagination approach the separation between clusters was considered. Recall that the proposed approach uses the Elbow method to derive a value for k to be used in the k-means clustering which is then validated by the Silhouette method. The evaluation was conducted with respect to

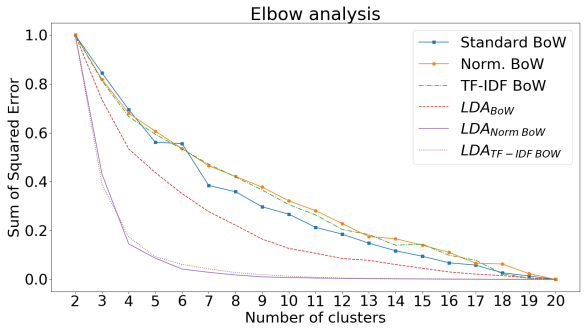

Figure (1) Elbow analysis of the Unspecified Label Pagination approach.

the six identified document representations. For the Elbow analysis the k-means clustering was performed for values of k ranging from 2 to 20 clusters. The resulting Elbow charts are presented in Figure 1. From the Figure an "Elbow point" can be identified at $k = 4$, particularly in the case where the LDA representation is built using a normalised bag-of-words. Analysis of silhouette coefficients obtained by clustering the bundles using this representation confirms $k = 4$ as the optimal cluster granularity.

Once we have the optimal number of clusters, each of the six BoW representations were clustered using k-means ($k = 4$). Ground truth data was exploited to compute the homogeneity, completeness and V-measure of the resulting clustering. A clustering result satisfies homogeneity if the clusters contain data points belonging to a single class. Completeness captures the capacity of the clustering to group together all the elements of one class. Both scores have values between 0.0 and 1.0. The V-measure, a conditional entropy-based external cluster evaluation measure, is defined as the harmonic mean of the homogeneity and completeness. A good clustering produces large values of both homogeneity and completeness. Table 3 shows the values of these metrics for the six BoW representations.

Multi-Dimensional Scaling (MDS) [6] was applied to visually analyse the obtained clusters. In general, the goal of this analysis is to detect meaningful dimensions that allow analysts to explain observed similarities or differences (distances) between the investigated clusters. However, MDS can be applied to reduce the number of features in the data with the aim of visually identifying which observations are similar.

Figure 2 shows the 2-dimensional cluster separation for the best clustering configuration obtained using the TF-IDF BoW representation. Each point in the figure represents a page within the document bundle. Clusters 1 and 2 in the figure represent the classes *"consent form"* and *"gp records"*, respectively. These two clusters can be easily identified. All the *"consent form"* data points were associated to the same cluster. Regarding the classes *"consultation note"* and *"blood test"* (Cluster 3 and 4 in Figure 2), 29 out of 30 data points were correctly associated to the correct cluster for both classes. This approach was least effective in clustering pages belonging to the *"gp records"* class. A total of seven *"gp records"* pages were assigned to the wrong cluster. In particular, four data points were assigned to the cluster containing *"blood test"* while the remaining three were included in the cluster containing consultation notes. This is due to

Representation	Random Forest											
	blood test			consent form			consultation note			gp record		
	Precis.	Recall	F1	Precis.	Recall	F1	Precis.	Recall	F1	Precis.	Recall	F1
Standard BoW	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**96.66%**	**98.3%**	96.77%	**100%**	98.36%
Norm. BoW	**100%**	96.66%	98.3%	**100%**	**100%**	**100%**	96.66%	**96.66%**	96.66%	96.77%	**100%**	98.36%
TF-IDF BoW	96.66%	96.66%	96.66%	**100%**	**100%**	**100%**	96.66%	**96.66%**	96.66%	**100%**	**100%**	**100%**
LDA_{BoW}	93.33%	93.33%	93.33%	96.66%	96.66%	96.66%	**100%**	**96.66%**	**98.3%**	90.32%	93.33%	91.8%
$LDA_{Norm\ BoW}$	73.33%	55%	62.85%	**100%**	**100%**	**100%**	41.93%	43.33%	42.62%	42.1%	26.66%	32.65%
$LDA_{TF-IDF\ BoW}$	45.83%	73.33%	56.41%	**100%**	96.66%	98.3%	38.88%	23.33%	29.16%	28%	23.33%	25.45%

Table (2) Classification performance (%) for the Random Forest classifier; results reported for each page category.

Representation	Completeness	Homogeneity	V-measure
Standard BoW	0.61	0.4	0.48
Norm. BoW	0.78	0.78	0.78
TF-IDF BoW	**0.81**	**0.81**	**0.81**
LDA_{BoW}	0.66	0.64	0.65
$LDA_{Norm\ BoW}$	0.78	0.41	0.54
$LDA_{TF-IDF\ BoW}$	0.71	0.43	0.54

Table (3) Clustering performance for the six BoW representations.

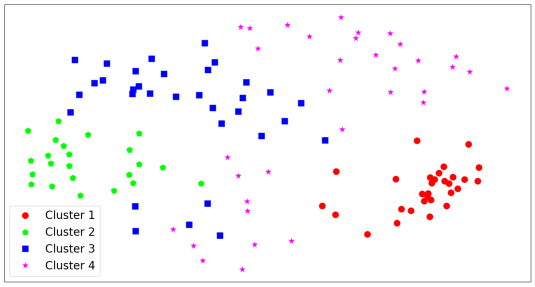

Figure (2) Visualisation, using 2-dimensional scaling, applied to the clustering obtained using the TF-IDF BoW representation.

the great variability of data included in *"gp records"*. In summary, the evaluation results obtained indicated that a process for selecting k by first determining its value using the Elbow technique, and then checking this using Silhouette Analysis, was appropriate.

6 CONCLUSIONS

Two mechanisms for automatically paginating bundles of legal documents have been presented. The first, fixed label pagination, adopted a supervised learning approach using a pre-defined set of labels and a training set. The second, unspecified label pagination, adopted an unsupervised learning approach. In both cases a range of document representations were considered for accident claim bundles of the form that might feature in accident claim litigation. In the case of the fixed label pagination approach, three alternative classifier generation models were used; it was found that Random Forest classifier generation and the standard BoWs document representation produced the best result. A best F1 score of 99.19% was recorded. In the case of unspecified label pagination, where cluster cohesion and separation were used as the performance metrics, it was found that the TF-IDF BoW produced the best result. The process allows the automatic extraction/parsing of useful information which speeds up case management. The proposed system is being tested by a group of legal professionals, and their initial feedback is positive. This testing is instructive in designing a system that will be maximally helpful for legal professionals operating in this domain. Ongoing work involves investigating how to filter pages according to their relevance to the case under examination.

REFERENCES

[1] Mehdi Allahyari, Seyed Amin Pouriyeh, Mehdi Assefi, Saied Safaei, Elizabeth D. Trippe, Juan B. Gutierrez, and Krys Kochut. 2017. A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques. *CoRR* abs/1707.02919 (2017). http://arxiv.org/abs/1707.02919

[2] Purnima Bholowalia and Arvind Kumar. 2014. Article: EBK-Means: A Clustering Technique based on Elbow Method and K-Means in WSN. *International Journal of Computer Applications* 105, 9 (November 2014), 17–24.

[3] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3, 4-5 (2003), 993âĂŞ1022.

[4] L. Karl Branting. 2017. Automating Judicial Document Analysis. In *Proceedings of the Second Workshop on Automated Semantic Analysis of Information in Legal Texts*.

[5] Leo Breiman. 2001. Random Forests. *Machine Learning* 45, 1 (01 Oct 2001), 5–32. https://doi.org/10.1023/A:1010933404324

[6] Andreas Buja, Deborah F. Swayne, Michael L. Littman, Nathaniel Dean, Heike Hofmann, and Lisha Chen. 2008. Data visualization with multidimensional scaling. *Journal of Computational and Graphical Statistics* (2008).

[7] Alexander Clark, Chris Fox, and Shalom Lappin. 2010. *The Handbook of Computational Linguistics and Natural Language Processing*. Wiley-Blackwell.

[8] Jack G. Conrad, Khalid Al-Kofahi, Ying Zhao, and George Karypis. 2005. Effective Document Clustering for Large Heterogeneous Law Firm Collections. In *Proceedings of the 10th International Conference on Artificial Intelligence and Law (ICAIL '05)*. ACM, 177–187.

[9] Corinna Cortes and Vladimir Vapnik. 1995. Support-Vector Networks. *Mach. Learn.* 20, 3 (Sept. 1995), 273–297.

[10] William M. Darling. 2011. A Theoretical and Practical Implementation Tutorial on Topic Modeling and Gibbs Sampling.

[11] M Ikonomakis, S Kotsiantis, and V Tampakas. 2005. Text classification using machine learning techniques. *WSEAS Transactions on Computers* 4, 8 (2005), 966–974.

[12] Mita K Dalal and Mukesh Zaveri. 2011. Automatic Text Classification: A Technical Review. *International Journal of Computer Applications* 28 (08 2011).

[13] David G. Kleinbaum and Mitchel Klein. 2010. Introduction to Logistic Regression. *Logistic Regression: A Self-Learning Text* (2010), 1–39.

[14] David MacKay. 2003. *Information Theory, Inference and Learning Algorithms*. Cambridge University Press.

[15] R. Mihalcea and P. Tarau. 2004. TextRank: Bringing Order into Texts. In *Proceedings of EMNLP-04 and the 2004 Conference on Empirical Methods in Natural Language Processing*.

[16] L. Page, S. Brin, R. Motwani, and T. Winograd. 1998. The PageRank citation ranking: Bringing order to the Web. In *Proceedings of the 7th International World Wide Web Conference*. 161–172.

[17] Juan Ramos. 2003. *Using TF-IDF to Determine Word Relevance in Document Queries*. Technical Report.

[18] Peter Rousseeuw. 1987. Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. *J. Comput. Appl. Math.* 20, 1 (Nov. 1987), 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

[19] Octavia-Maria Sulea, Marcos Zampieri, Shervin Malmasi, Mihaela Vela, Liviu P. Dinu, and Josef van Genabith. 2017. Exploring the Use of Text Classification in the Legal Domain. *CoRR* abs/1710.09306 (2017).

[20] Ravi Kumar V and K. Raghuveer. 2012. Article: Legal Documents Clustering using Latent Dirichlet Allocation. *International Journal of Applied Information Systems* 2, 6 (2012), 27–33.

[21] H. Zhang and D. Li. 2007. NaÃŕve Bayes Text Classifier. In *2007 IEEE International Conference on Granular Computing (GRC 2007)*. 708–708.